빅데이터가 만들어 내는 본질적인 변화에 대한 고찰

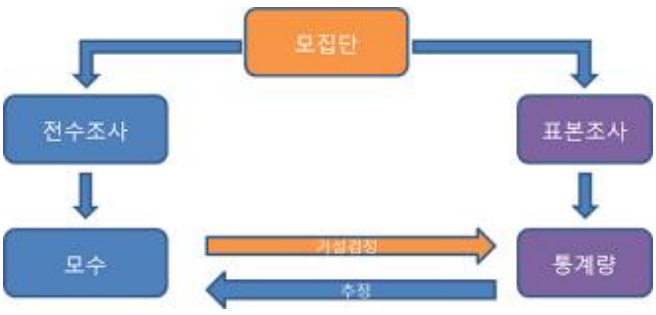
한신대학교 응용통계학과 4학년 201452024 박상희

빅데이터라는 말은 정확하게 언제부터 어디서부터 시작되었을까? 빅데이터는 어느 날 갑자기 생겨난 것이 아니다. CRM에서 데이터 마이닝, 데이터 마이닝에서 빅데이터로 우리가 기존에 알고 있던 데이터를 다루는 방법이 데이터의 크기가 확장되고 처리 기법이 바뀌면서 이름만 점점 바뀌어 왔던 것이다. 그렇다면 데이터 마이닝에서 빅데이터로 도대체 무엇이 바뀌었길래 얼마나 큰 변화가 있었길래 빅데이터라는 새로운 용어가 등자하고 4차 산업혁명이라는 단어가 등장하게 되었을까? 다음 4가지는 빅데이터가 만들어낸 우리 시대의 본질적인 변화이다. 그러나 아직은 완벽하게 변화되지 않았으며 부작용도 나타나고 있다.

▣ 사전 처리에서 사후처리 시대로

실험계획법(Experiments Design)이란 연구자가 가설을 설정하고 그 가설을 입증하기 위한 데이터를 설계하고 수집하는 통계적 분석 방법이다. 일반적으로 우리는 데이터를 수집할 때 우리가 필요한 정보, 우리의 계획에 필요한 정보만 사전처리하여 수집하곤 하였다. 그러나 빅데이터 시대에 오면서 사후처리로 바뀌기 시작하였다. 일단 데이터를 모은 후 그 다음 설계자가 그 안에서 새로운 데이터를 생성하기도 하고, 필요 없는 데이터는 버리기도 하는 방식이다. 사전 처리 방법에서는 설령 필요한 데이터가 숨겨져 있더라도 한 번 제거를 하면 돌이킬 수 없지만, 사후처리는 원데이터를 보존하고 있기 때문에 숨겨진 유용한 데이터도 다시 가공할 수 있다는 장점이 있다. 그렇기 때문에 대용량 데이터를 가공하는 기술도 분석가가 갖춰야할 능력이 되었다. 이전의 분석가들은 단순히 가지고 있는 데이터만 잘 분석하면 되었지만, 지금의 빅데이터 분석가들은 데이터를 잘 핸들링(Handling)하는 기술도 많이 요구되어지고 있다.

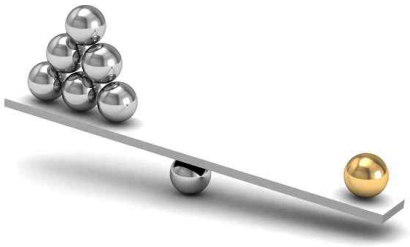
▣ 표본조사에서 전수조사로



통계조사의 방법에는 전수조사와 표본조사가 있다. 전수조사는 내가 조사하고자 하는 대상(모집단)의 특성을 직접 조사하는 것이고, 표본조사는 내가 조사하고자 하는 대상(모집단)을 현실적으로 조사하기 힘들고 비용이 많이 들어 확률추출에 기반해 표본을 뽑고 그 표본을 통해 모집단의 특성을 추정하는 방법이다. 그러나 빅데이터의 발전으로 기업이나 기관 등에서 방대한 데이터를 전수조사할 수 있는 기술여건이 마련되었다.

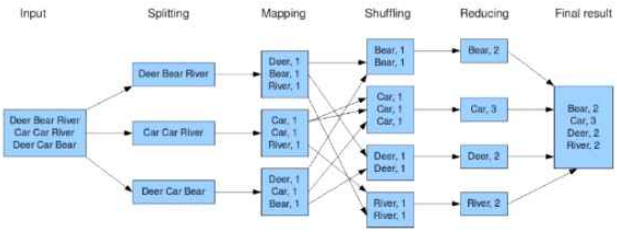
예를들어 서울에 사는 20대 남성의 해외 여행수를 알고 싶다면, 예전에는 표본조사를 실시하여 모집단을 추정하였지만, 지금은 국내 주요 통신사들과 협력하여 거주지가 서울인 있는 20대 남성의 데이터를 받은 뒤 공항의 데이터베이스와 연결하여 대상자들의 여행 기록을 보면 알 수 있다. 이처럼 빅데이터는 여러 분야에서 현실적으로 불가능할 것만 같았던 전수조사를 가능하게 했지만, 한편으로는 개개인의 사생활 침해, 개인 정보 유출 등 아직 해결해야 할 문제가 많아 쉽게 사용하지는 못한다. 개개인의 정보를 단순히 데이터로만 생각하기 보다는 좀 더 고민하고 신중하게 사용해야 한다.

▣ 질보단 양으로



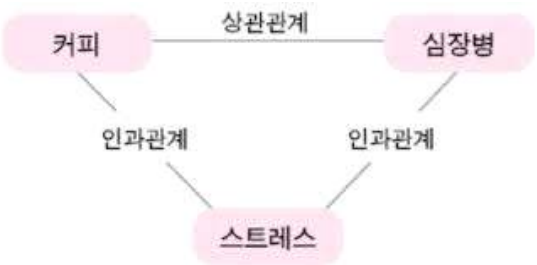
통계학자들에게 “변수가 100개이다.” 라는 말을 던졌을 때 대부분의 반응들은 “변수가 아니라 관측치를 잘못 말 한 것이다.” 라고들 할 것이다. 왜냐하면 일반적으로 데이터를 수집할 때 변수들은 보통 10개 이하로 수집되기 때문이다. 실험을 설계하고 변수 하나를 선택하고 측정하는 것에도 엄청난 돈과 시간이 들어가기 때문이다. 그러나 빅데이터는 이런 식으로 설계를 하여 데이터를 수집하는 것이 아니다. 설계가 없이 자연스럽게 만들어진 데이터가 대부분이기 때문이다. 그렇기 때문에 빅데이터를 가지고 분석을 할 경우에는 기존의 변수들이 의미가 없을 가능성이 매우 높다. 따라서 분석가가 이런 변수들 사이에서 유의미한 새로운 변수를 생성해 내기도 하고, 분석에 필요한 변수들을 선택하기도 하며 변수들을 가공하는 작업이 매우 중요하게 되었다.

또한 이렇게 대용량 데이터들을 처리하기 위한 기술도 매우 중요해졌다. 맵리듀스(MapReduce)란 이런 빅데이터들을 빠르고 정확하게 처리하는 기술이다. 맵(Map)은 여러 개의 task에 각자 데이터들을 할당하고 연산을 수행한 뒤 리듀스(Reduce)라는 작업을 통해서 하나의 결과를 만들어내는 기술이다.



이 맵리듀스의 핵심은 “병렬처리 시스템” 이다. 이전까지는 컴퓨터 1대가 모든 연산을 수행하고 결과까지 계산한 반면, 병렬처리 시스템은 여려 대의 계산 컴퓨터가 있으며 연산 결과만 따로 계산하는 컴퓨터 역시 존재하여 대용량의 데이터를 처리하는 속도를 매우 단축시켰다. 따라서 빅데이터 시대에는 단순히 데이터를 분석하는 분석가뿐만 아니라 이런 데이터들을 처리하고, 데이터들의 흐름을 제어하는 분야도 떠오르는 산업이다.

▣ 인과관계에서 상관관계로



여러 연구에서 대학 졸업 여부와 소득 사이에는 상관관계가 존재한다고 밝혔다. 즉. 대학 졸업자들 중 대부분은 소득이 높은 경향을 보이는 것으로 나타났다. 하지만 이는 인과관계를 의미하지는 않는다. 예를 들어 대학을 나와야 소득이 높아진다고 해석해서는 안 된다. 실제로 소득이 높아서 대학에 진학했을 수도 대학을 안 나와도 높은 소득을 얻을 수 있기 때문이다. 우리가 어떤 관계를 분석할 때 상관관계와 인과관계를 잘 구분해야 하는 이유이다.

그러나 실제로 상관관계와 인과관계를 구분하는 것은 쉽지 않다. 상관관계는 단순히 데이터에서 패턴을 찾아내는 것으로 데이터만 주어진다면 관계를 찾아낼 수 있지만, 인과관계는 어떠한 이론이나 검증된 모델을 바탕으로 관계를 추론하는 것이기 때문이다. 그렇다면 단순히 패턴만 찾아내는 상관관계는 중요하지 않는 것일까? 만약 예측이 목적이라면 패턴은 매우 유용한 정보를 제공할 것이다. 그러나 근본적인 원인을 밝혀내고 그에 대한 대책을 마련하는 것이 목표라면 모델을 세워 인과관계를 찾는 것이 바람직하다.

데이터는 말을 하지 않는다. 또한 항상 언제나 그대로이다. 데이터를 다루고 변형시키고 하는 것은 그것을 다루는 사람이다. 우리가 무엇을 위해 이 데이터를 사용하는지 정확하게 인지하고 있어야 데이터가 우리의 목적을 달성하기 위한 도움이 될 것이다.